

Hema Harsha Vardhan Peela

Data Engineer

hemaharshavardhanpeela@gmail.com | +1 8065440868 | [LinkedIn](#) | [GitHub](#) | [Portfolio](#)

SUMMARY

Driven Data Engineer (F1 Student) with hands-on experience designing and deploying scalable, reliable data pipelines on cloud platforms. Proven ability to optimize query performance by 40% and reduce quality-related escalations by 25%. Eager to leverage strong Python, Spark, and Airflow expertise in ETL processes and data modeling to transform complex data into trustworthy datasets that power AI-driven platforms, enabling critical insights and data-driven decision-making.

Data & Cloud Engineer | YourBook Team | USA | CPT

Nov '24 – May '25

- Redesigned cloud warehouse schemas (PostgreSQL, Redshift), optimizing partitioning and joins to boost query performance by 40%
- Developed and maintained systems supporting the Analytics Infrastructure & Data Lake, focusing on data pipeline development and operations.
- Collaborated with cross-functional teams to integrate operational and user behavior data into centralized storage systems
- Collaborated on performance tuning at code and infrastructure levels, integrating diverse data into centralized storage.
- Built Power BI/SQL dashboards to monitor warehouse performance and fulfillment accuracy, reducing quality escalations by 25%.
- Assisted in data transformation/onboarding and participated in incident response, improving data consistency and system reliability

Data Engineer | Xpress Global | India

Apr '21 – May '23

- Designed, built, and maintained robust ETL/ELT pipelines to ingest, transform, and load large datasets from diverse sources (e.g., customer behavior logs, internal databases) into centralized data warehouses/data lakes.
- Optimized data models and database schemas (e.g., PostgreSQL, Redshift) to improve query performance and data accessibility for downstream analytics and reporting.
- Developed and deployed scalable data processing jobs using frameworks like Spark (PySpark) or cloud-native services (AWS Glue, Lambda) to handle high-volume data streams

PROJECTS

Sales Performance Optimization Using Data Pipeline Automation | AWS, Python, SQL, Tableau

Mar '25

- Designed and deployed a scalable ETL pipeline using Python, AWS Glue, and Lambda to ingest and standardize raw sales data from multiple regions, making it analysis-ready for business intelligence applications
- Implemented scalable data transformation logic using Azure Databricks notebooks (PySpark) and performed performance tuning to optimize data processing speeds.
- Implemented sales trend analysis using time-series forecasting in Python, improving forecast accuracy by 25% and aiding revenue planning.
- Developed interactive Tableau dashboards for sales leadership to track KPIs such as conversion rates and customer retention. to derive strategic insights aligned with business growth and revenue optimization.

Garbage Data Filtering in Social Networking Sites | Pyspark, PyArrow, Scikit-learn, pandas, Matplotlib

Jan '25

- Analyzed social media platform data to identify behavioral patterns and anomalous activity, developing a model to filter spam and improve content quality by 45%.
- Spearheaded the implementation of the state-of-the-art algorithms, including 'Random Forest,' 'XG Boost,' 'Decision Tree,' and 'Spark Naive Bayes,' elevating the precision and classification capabilities of the machine learning model by 25%.

Customer Segmentation using Machine Learning in R | GridExtra, Plotrix, NBClust, FactoExtra, Matplotlib

Oct '24

- Implemented a highly innovative Machine Learning Model that categorizes customers, leading to a 30% reduction in customer churn and a 20% increase in customer lifetime value.
- Led the strategic development and execution of a sophisticated K-Means Clustering model, elevating customer satisfaction and loyalty through precise customer classification; the implementation achieved a 40% increase in targeted marketing effectiveness.

Student Result Analysis | ReactJs, Django, Python

Aug '24

- Built a scalable ETL pipeline using AWS Glue and Lambda to process large-scale financial and behavioral CSV data into SQL tables, enabling automated analysis and reducing manual effort by 80%.
- Developed a ReactJS-Django platform with backend APIs and interactive filters to visualize student performance; applied statistical models and dashboards to cut reporting time by 20% and improve academic planning.

TECHNICAL SKILLS

Programming Languages: Python, Java, C++

Databases: PostgreSQL, Redshift, MySQL, SQL Server, MongoDB, GCP BigQuery

Modeling: Star Schema, Snowflake Schema

Big Data/ETL: Apache Spark (PySpark), AWS Glue, GCP Dataflow, Airflow, Kafka

Machine Learning: KMeans, Random Forest, XGBoost, Naive Bayes, A/B Testing, Time-Series Forecasting

Visualization Tools: Tableau, Power BI, Looker, Google Sheets, Excel, PowerPoint, Google Slides

Cloud/Data: GCP (BigQuery, PubSub, DataProc), AWS (Redshift, Glue, Lambda)

Libraries & Frameworks: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn

Other Technologies: Docker, Git, GitHub

EDUCATION

Masters in Computer Information Science – Texas Tech University | 3.7/4.0

Aug '23 - May '25

Relevant Coursework: Analysis of Algorithms, Software Modelling and Architecture, Advanced Operating Systems, Neural Networks, Computer Organization and Architecture, Advanced Database Management.

Course Awards: TTU Distinguished Student Award.